

Data Science Consortium: Access to Scientific Data Anywhere by Anyone
January, 2012

Whitepaper Outcome of NSF DataNet/INTEROP (Data2012) meeting
January 25-27, 2012, Indianapolis, IN

Data that contributes to and advances the science, engineering, and technology research mission is growing in size at a pace not seen heretofore. It has been suggested that a 4th paradigm of science has emerged by which discovery is carried out primarily through analysis of data. Universities, funding agencies, national organizations, and researchers themselves are grappling with ways in which scientific data can be preserved and effectively shared today and into the future. There have been numerous meetings, workshops, and reports defining and discussing the problem and its many aspects; it is now time for concrete action.

In January 2012 NSF sponsored a workshop held in Indianapolis that brought together 70 or so participants including the NSF DataNet projects and INTEROP projects as well as representatives from national and international organizations. From the workshop emerged the idea of an open organization around scientific data management that fosters agreement and approaches to managing scientific data through the data lifecycle with emphasis on access, sharing, preservation, governance and use. The workshop devoted several hours to discussion of the forms such an organization could take. This whitepaper captures the development of the idea over the day and a half of discussion.

It was firmly agreed that the organization in whatever form must be open to all. In acknowledgement of the similar problems worldwide, the organization must have strong ties internationally. The organization's primary objective is to produce high quality, relevant technical documents that influence the way people manage, access, govern, and store scientific data. These documents include protocol standards, service definitions, data models, query languages, best current practices, and information documents of various kinds.

Those likely to participate in the organization include digital library and information scientists, computer scientists, domain scientists seeking answers, policy experts, program funding officers, data center managers, and industry members. An approach to progress, as embodied very successfully in IETF, is rough consensus and running code.

The discussion identified two ways of moving forward, and a recommendation that these be pursued in parallel. The first is through establishment of a small group that explores the organization development more fully to clarify its vision and charge. The group also explores synergies and alignments with other emerging and established organizations (DAITF in Europe, Digital Preservation Network (DPN) being led by university CIOs and libraries, ESIP, W3C) who have complementary

and/or overlapping goals. Several options for realization of the organization through the IETF governance framework were discussed.

The second complementary way of moving forward is through establishment of a working group that identifies and organizes a couple of first step, concrete technical activities to advance the community in a pragmatic and practical way.

- One suggested activity is setting up and running a challenge. A challenge topic is chosen such that it identifies a common pain point, a technical problem of interest to multiple participants. The challenge could be carried out over a several month period. The series of provenance challenges [] are an example. A challenge topic put forth is high level discovery of scientific data through a Google-type search. This could be followed by a data retrieval and visualization challenge.
- A shorter-term hackathon could have the community converge for an intense a 2-3 day effort to focus attention on a shared problem. These small projects are chosen because they a) involve some important elements of (or are closely related to) a person's current research interests, b) can be completed or significantly moved forward within the challenge period, and c) involve collaboration between a group of interested people drawn from multiple venues or projects. Towards the end of the challenge period, the working groups report back to one another and share lessons.

Next steps. Several attendees of the DataNet/INTEROP meeting will attend the DAITF meeting in Copenhagen March 2012. There are plans to convene a small group to attend IETF in Vancouver July 2012 to more formally explore relations through the IETF organization. NSF funding through the DataNet and INTEROP programs has created and brought together a powerful element of technology contributors and researchers who deeply understand and are trying to advance the scientific data management problem in its world proportions. A collective effort will allow synergies, best practices, and standards to emerge. The time is right.

To join the discussion, receive progress reports, review notes from convened meetings and see who else is involved visit <http://d2i.indiana.edu/data2012/> for more information.